NBER WORKING PAPER SERIES

IMPROVING SCHOOL ACCOUNTABILITY MEASURES

Thomas J. Kane
Douglas O. Staiger

Working Paper 8156
http://www.nber.org/papers/w8156

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2001

We thank Terry Moe of the Hoover Institution at Stanford for providing tabulations of the 1993-94 Schools and Staffing Survey. Seminar participants at University of California-Berkeley, Stanford University, Columbia University and Harvard University provided a number of helpful suggestions. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Improving School Accountability Measures
Thomas J. Kane and Douglas O. Staiger
NBER Working Paper No. 8156
March 2001
JEL No. I2

ABSTRACT

A growing number of states are using annual school-level test scores as part of their school accountability systems. We highlight an under-appreciated weakness of that approach – the imprecision of school-level test score means -- and propose a method for better discerning signal from noise in annual school report cards. For an elementary school of average size in North Carolina, we estimate that 28 percent of the variance in 5th grade reading scores is due to sampling variation and about 10 percent is due to other non-persistent sources. More troubling, we estimate that less than half of the variance in the mean gain in reading performance between 4th and 5th grade is due to persistent differences between schools. We use these estimates of the variance components in an empirical Bayes framework to generate "filtered" predictions of school performance, which have much greater predictive value than the mean for a single year. We also identify evidence of within-school heterogeneity in classroom level gains, which suggests the importance of teacher effects.

Thomas J. Kane
Kennedy School of Government
79 JFK Street
Cambridge, MA 02138
tom_kane@harvard.edu

Douglas O. Staiger
Department of Economics
Dartmouth College
Hanover, NH 03755
douglas.o.staiger@dartmouth.edu

I. Introduction

By the spring of 2000, forty states had begun using student test scores to rate school performance. Moreover, a growing number of states are attaching explicit monetary rewards or sanctions to a school's test performance. For example, California plans to spend $677 million on teacher incentives this year, providing bonuses of up to $25,000 to teachers in schools with the largest test score gains. In this paper, we highlight an under-appreciated weakness of school accountability systems-- the imprecision of test score measures in identifying long-term differences between schools in student performance-- and propose a method for better discerning the signal amidst the considerable noise of annual school report cards.

The imprecision of test score measures arises from two sources. The first is sampling variation, which is a particularly striking problem in elementary schools. With the average elementary school containing only 60 students per grade level, the amount of variation due to the idiosyncracies of the particular sample of students being tested is often large relative to the total amount of variation in student performance observed. A second source of imprecision arises from one-time factors that are not sensitive to the size of the sample: a dog barking in the playground on the day of the test, a severe flu season, one particularly disruptive student in a class or favorable "chemistry" between a group of students and their teacher. Both small samples and other one-time factors that are not sensitive to sample size can add considerable volatility to test score measures.

At first perusal, one might be surprised that there would be such volatility at the school level, since one would expect the averaging of student scores to dampen such noise. However, it is not the absolute amount of imprecision that matters, but the amount of imprecision relative to

the underlying signal that dictates the reliability of any particular measure. Although the averaging of students' scores does help dampen volatility, even small fluctuations in a school's score can have a large impact on a school's ranking, simply because schools' test scores do not differ dramatically in the first place. This reflects the longstanding finding from the Coleman report (*Equality of Educational Opportunity*, issued in 1966), that less than 15 percent of the variance in student test scores is between-schools rather than within-schools.[1] We estimate that the confidence interval for the average 5th grade reading score in a school with 60 students per grade level would extend from roughly the 25th to the 75th percentile.[2]

Such volatility can wreak havoc in school accountability systems. First, to the extent that test scores bring rewards or sanctions, school personnel are subjected to substantial risk of being punished or rewarded for results beyond their control. Moreover, to the extent such rankings are used to identify best practice in education, virtually every educational philosophy is likely to be endorsed eventually. For example, when the 1998-99 MCAS test scores were released in Massachusetts in November of 1999, the Provincetown district showed the greatest improvement over the previous year. The *Boston Globe* published an extensive story describing the various ways in which Provincetown had changed educational strategies between 1998 and 1999, interviewing the high school principal and several teachers.[3] As it turned out, they had

---

[1]Coleman *et. al.* (1966). Table 3.22.1 in the report suggest that 7 to 14 percent of the variance in verbal achievement in 1st, 3rd, 6th, 9th and 12th grades is "between-schools" rather than "within-schools".

[2]This includes the volatility due to sampling variation. Any volatility due to other non-persistent variation would widen the range of potential scores.

[3]Brian Tarcy, "Town's Scores the Most Improved" *Boston Globe*, December 8, 1999, p. C2.

changed a few policies at the school-- decisions that seemed to be validated by the improvement in performance. One had to dig a bit deeper to note that the Provincetown high school had only 26 students taking the test in 10th grade. Given the wide distribution of test scores among students in Massachusetts, any grouping of 26 students is likely to yield dramatic swings in test scores from year to year, that is large relative to the distribution of between-school differences. In other words, if school-level test scores are the gauge, the *Boston Globe* and similar newspapers around the country will eventually write similar stories praising virtually every variant of educational practice. It is no wonder that the public and policymakers are only more confused about the way to proceed.

In this paper, based on methods developed in McClellan and Staiger (1999) for the analysis of hospital performance measures, we decompose the variation across schools and over time in test score measures into three components: that which reflects persistent differences in performance between schools, sampling variation, and other non-persistent differences among schools.

Decomposing the variation in school test scores into each of these components is useful for two purposes. First, it provides insight into questions such as: Are there sizeable persistent differences in student outcomes across schools? How correlated are these differences over time, across grades, or across subject tests? And what is the signal-to-noise ratio of school performance measures currently in use? Our second purpose is to use the resulting information to construct more accurate predictions of school performance. In particular, we use the estimated variance components in an empirical Bayes framework to construct "filtered" estimates of each school's performance measures. These filtered estimates optimally incorporate information from past

years and other grades or subject tests in order to eliminate much of the volatility observed in conventional measures and better predict school performance.

Using math and reading test scores for 3rd, 4th and 5th grade students in North Carolina from 1993 through 1999, we organize the results in the paper into four broad areas: First, we decompose the variation in 5th grade math and reading scores (measured as both test score levels and test score gains from the previous year) into sampling variation, a persistent component of the signal, and a non-persistent component of the signal. For both levels and gain scores, we find that sampling variation accounts for roughly 15 percent of the variance across schools in 5th grade math scores and 30 percent of the variation in 5th grade reading scores for an average size elementary school in North Carolina, and an even larger share in smaller schools.

Reading scores contain a larger share of non-persistent "noise" than math scores. The reason is that there is simply less signal variation in reading test scores, not because higher absolute levels of sampling variation. For example, we estimate that the standard deviation across schools in the persistent component of reading scores is about half that of math scores. Thus, reading scores are not particularly good indicators of school performance.

The proportion of variance attributable to persistent and non-persistent components varies dramatically between test score levels and gains. For example, the non-persistent component accounts for 8 percent of the variance between schools in math test score levels in 5th grade and 29 percent of the variance in gains. Although gain scores are often touted as better indicators of a school's "value-added", they are much more likely to be affected by idiosyncratic fluctuations in scores from year to year. (The impact of one-time events-- such as the weather on the test date-- are multiplied by two when moving from test score levels to test score gains.) Moreover,

schools differ much less in their average gains than in their average test score levels. In other words, attempting to estimate a school's value-added is analogous to looking for a smaller needle in a bigger haystack.

The second focus of our analysis is on evaluating the accuracy of "filtered" estimates, as compared to more naive predictions of school performance that use only the most recent year's data or an average over recent years chosen on an ad hoc basis. Rather than *arbitrarily* choosing to portray a given school's performance with the average score from the latest year or the latest two years or the latest five years, the filtered estimates use past relationships to optimally construct a summary of all past performance measures with the greatest predictive power for current levels of performance. That summary measure for math performance, for instance, includes a weighted average of recent math scores as well as reading scores. Moreover, the weights vary by school size to reflect the greater sampling variation in the recent data from smaller schools. We find that the filtered estimates are far more accurate than more naive estimates of school performance. For example, filtered estimates of school performance from 1997 generate an out-of-sample forecast R-squared of .27 and .31 for math and reading scores in 1999, while using the 1997 value of each score as the forecast generates a negative forecast R^2 for both math and reading scores in 1999. (A negative forecast R^2 simply means that one would have achieved a lower mean-squared error by discarding each school's score and simply using the statewide average to predict each school's score.)

The third focus of our analysis is on the relationship between test score improvements and other measures of student inputs. While it is difficult to directly test whether greater student familiarity with the test or teachers' "teaching the test" is leading to inappropriately inflated test

scores, we investigate whether the schools with substantial increases in student test score gains between 1994 and 1999 also experienced improvements in student engagement, as indicated by student absences, the amount of homework students report doing or the amount of TV they report watching. Although school differences in TV watching and homework were strongly related to differences in test score performance in the baseline year (1994), there is little evidence that the schools that improved their test performance between 1994 and 1999 also improved on any of these other measures of student engagement.

The final focus of our analysis is on exploring the importance of classroom (e.g. teacher), as opposed to school effects on student performance. We find that, once one accounts for the variance in student test scores due to sampling variation and other non-persistent factors, the pattern of test performance that emerges hints at substantial heterogeneity in gains in different classrooms within the same school. Evidence from the one year of data for which we can identify students in individual classrooms suggests that the within-school variance in student performance across classrooms is roughly equivalent to the between school variance in student performance. Somewhat more indirectly, we find that there is less variation in student performance across schools with many classrooms, as one would expect if the law of averages reduced the impact of classroom variation on the large school's performance. We also find there is much more correlation in same-grade math gains from one year to the next in the same school than there is between 4th and 5th grade math gains in the same year at the same school. Moreover, the declining pattern of correlation over time is broadly consistent with the teacher turnover rate. Similarly, while a school's mean math and reading gains are highly correlated in 4th and 5th grades (with a correlation coefficient of .7 to .8) where a single teacher teaches all

subjects, the correlations are weaker in 7th and 8th grades when teachers specialize in different subjects (correlation of .5 to .6).

II. The North Carolina Test Score Data

We obtained math and reading test scores for nearly 300,000 students in grades 3 through 5, attending one of roughly 1300 elementary schools each year between the 1992-93 and 1998-99 school years. (The data were obtained from the N.C. Department of Public Instruction.) Although the file we received had been stripped of student identification numbers, we were able to match a student's test score in one year to their test score in the previous year using data on day of birth, race and gender.[4] In 1999, 84 percent of the sample had unique combinations of birth date, school and gender. Another 14 percent shared their birth date and gender with at most 1 other students in their school and grade and 2 percent shared their birth date with 2 other people. (Less than 1 percent shared their birth date and gender with 3 or more students in the school and no match was attempted for these students.) Students were matched across years only if they reported the same race. If there was more than 1 person with the same school, birth date, race and gender, we looked to see whether there were any unique matches on parental education. If there was more than one person that matched on all traits-- school, birth date, race, gender and parental education-- -the matches that minimized the squared changes in student test scores were kept.

[4]In addition, the survey contained information on parental educational attainment reported by students. Given changes in student responses over time, we did not use parental education to match students from year to year, although we did use these data to when attempting to control for the socioeconomic background of students.

However, because of student mobility between schools or student retention, the matching process was not perfect. We were able to calculate test score gains for 65.8 percent of the 4th and 5th grade students in 1999. (The matching rate was very similar in other years.) Table 1 compares the characteristics of the matched and the non-matched sample of fourth and fifth grade students in 1999. The matched sample had slightly higher test scores (roughly .2 student level standard deviations in reading and math), a slightly higher proportion female, a slightly higher proportion black and Hispanic, and a slightly lower average parental education than the sample for which no match could be found.

III. Volatility in Student Test Scores

In the analysis in this paper, individual student test scores and test score gains were first regression-adjusted, including dummy variables parent's education, gender and race/ethnicity as well as fixed effects for each school and year. Figure 1 portrays the distribution of regression-adjusted mean math and reading 5th grade gain scores for each school by the number of students in each grade. (Test scores have been reported in standard deviations of student-level math test scores. The mean gain is the average change in students' test scores between the end of grade 4 and the end of grade 5.) Two facts are immediately apparent: First, virtually all of the schools with the highest mean test scores or mean gain scores (for example, schools with mean test scores more than one student-level standard deviation above the mean) were small schools, with fewer than 40 students per grade level. However, there was little difference in mean performance by school size. Indeed, the poorest performing schools were also primarily small

schools. Second, this variability in test scores among small schools is not solely due to heterogeneity among small schools. The graphs on the right in Figure 1 report the change between calendar years in each school's mean test score and mean gain score. The small schools are also much more likely to report large *changes* in mean scores and mean gains from one year to the next, both positive and negative.

Although we are using data from North Carolina, the issue is not unique to North Carolina. Figure 2 reports average 4th Grade Math Scores in 1999 and changes in math scores between 1999 and 2000 in California by school size. (These data were not regression-adjusted, because we did not have the student-level data.) There is a similar pattern of wider dispersion among smaller schools, particularly in changes in mean scores from 1999 to 2000. California will be basing its incentive rewards of up to $25,000 per teacher on changes in annual test scores at the school level.

Table 2 provides another illustration of the volatility in school test score measures. Each year between 1994 and 1999, we ranked schools by their average test score levels and average test score gains in 5th grade, after adjusting for race, parental education and gender (as in specifications (b) and (d) above). We counted the proportion of times each school ranked in the top 10 percent over the six-year period. If there were "good" and "bad" schools which could be observed with certainty, we might expect to see 90 percent of schools never ranking in top 10 percent and 10 percent of schools always ranking at the top. At the opposite extreme, where schools were equal and the top 10 percent were chosen by lottery each year, we would expect 47 percent schools ranking in the top 10 percent at least once over 6 years and only 1 in a million ranking in the top 10 percent all 6 years.

The rankings generally resemble a lottery, particularly in gain scores. If math scores were the metric, between 31 and 36 percent of schools would have ranked in the top 10 percent at some point over the 6 years, depending upon whether one used the mean test score or the mean gain in test scores. Less than 1 percent of schools ranked in the top 10 percent all 6 years. Reading test scores seem to have an even larger random component, with 35 to 38 percent of schools ranking in the top 10 percent at some point over 6 years and less than one percent of schools ranking in the top 10 percent for all 6 years. No school ranked in the top 10 percent on 5th grade reading gains for all 6 years.

Small sample size is a particularly large problem for elementary schools. However, the problem is not unique to elementary schools. Figure 3 portrays the distribution of sample sizes by grade in North Carolina. School size is generally smaller in grade 4. However, while the size of the average middle school is larger than the size of the average elementary school, there is also more heterogeneity in school size. The same phenomenon is exaggerated at grade 9. In other words, elementary schools tend to be smaller than middle schools and high schools. However, they are also more uniform in size, meaning that schools have a more similar likelihood of having an extremely high or extremely low score due to sampling variation. Middle schools and high schools have larger sample sizes on average, but there is greater heterogeneity in the likelihood of seeing an extremely high or extremely low test score due to sampling variation.

IV. A Method for Filtering School Performance Data

To dampen volatility, accountability systems tend to gravitate toward higher levels of aggregation-- raising units of measurement from the classroom level, to grade levels, to whole schools, to school districts. Yet, at higher levels of aggregation, any individual teacher or principal could expect to have only a miniscule impact on the overall score. When test scores are based upon average performance of hundreds of teachers and thousands of students, the notion of holding individual students or teachers "accountable" for their performance becomes vacuous.

Rather than aggregating across organizational units, our goal is to improve methods for aggregating at the school or classroom level across time, developing a method which optimally weights performance in recent years to generate an estimate of current levels of performance. Thus, to better estimate a school's performance on 5th grade math in 1999, our method uses all relevant information available from previous years, other grades, and other subject tests. More specifically, we employ an empirical Bayes estimator (Morris, 1983) similar to that developed by McClellan and Staiger (1999) to evaluate hospital performance. The estimation proceeds in two steps. The first step uses GMM methods to decompose the variance and covariance of the observed performance measures over time into signal and noise (e.g. estimation error) components, with the signal component being further decomposed into persistent (over time) and non-persistent components. The estimates of these variance components are of direct interest, as they provide information that is relevant to many substantive debates about the determinants of student performance. The second step uses the information on the variance components to form optimal linear predictions (or forecasts) of each performance measure (or any of its components) as a function of all the observed performance measures. We refer to these predictions as

"filtered" estimates, since the key advantage of such estimates is that they optimally filter out the noise in the observed performance measures.

The filtered estimates have a number of attractive properties. First, they incorporate information from many performance measures and many years in a systematic way into the predictions of any one performance measure, rather than relying on ad hoc averaging across measures or years. In particular, the filtered estimates are optimal linear predictions in the sense that they minimize the mean squared error of the prediction and thereby reduce noise in these measures to the maximum extent possible. A second property of these predictions is that, under assumptions of normality, they provide estimates of the posterior mean (and distribution) of each performance measure, conditional on all available data. If individual utility depends on these performance measures, then the posterior distribution contains all of the relevant information for the purpose of comparing schools based on expected utility. A third property of the filtered estimates is that regression coefficients are not attenuated when these estimates are used as independent variables (see Hyslop and Imbens, 2000). In contrast, coefficient estimates are attenuated towards zero when using conventional performance measures as independent variables because of the classical measurement error in conventional estimates. A final property of our filtered estimates is that they are quite easy to construct, with estimation taking only a few minutes on an average personal computer in our sample of roughly 1000 schools observed over 6 years with over 50,000 students per grade in each year. In contrast, estimation complexity has limited the application of most existing Bayesian approaches to relatively small samples (e.g. Normand, Glickman, and Gatsonis, 1997).

Many of the key ideas behind the filtered estimates are illustrated through a simple

example. Suppose that a school administrator is attempting to evaluate a particular school's performance based on the mean test scores of the students from that school in the most recent two years. For simplicity, suppose that these scores are standardized so that zero represents the state average and greater (less) than zero indicates a school that is above (below) the state average. Consider the following three possible approaches: (1) use only the most recent score, (2) construct a simple average of the scores from the two recent years, and (3) ignore the school's scores and assume that student performance in the school is equal to the state average. The best choice among these three approaches depends on two important considerations: the signal-to-noise ratio in the school's data, and how strongly correlated performance is from one year to the next. For example, if the average test scores for the school were based on only a few dozen students, and one had reason to believe that school performance did not vary much across the state, then one would be tempted to choose the last option and place less weight on the school's scores because of their low signal-to-noise ratio. Alternatively, if one had reason to believe that school performance changed very slowly over time, one might choose the second option in hopes that averaging the data over two years would reduce the noise in the estimates by effectively increasing the sample size in the school. Even with large samples of students being tested one might want to average over years if idiosyncratic factors such as the weather on the day of the test affected scores from any single year. Finally, one would tend to choose the first option, and rely solely on scores from the most recent year, if such idiosyncratic factors were unimportant, if the school's estimate was based on a very large sample of students, and if performance was likely to have changed from the previous year.

Our method of creating filtered estimates formalizes the intuition from this simple

example. The filtered estimates are a combination of the school's own test score, the state average, and the school's test scores from past years, other grades, or other subjects. As suggested by the example, to form the optimal combination one must know the amount of noise and signal variance in each measure, as well as the correlation across measures in the noise and signal variance. We estimate the noise variance (and covariance) in a straightforward manner for each school, based on the number of students being tested in each year. To estimate the signal variance (and covariance) for each performance measure, we subtract the noise variance from the total variance observed in each measure across schools (which reflects both signal and noise variance). In other words, we do not directly observe the "signal" of school quality and cannot measure its variance directly -- so instead we estimate the total variance in our imperfect measures of school quality, and then subtract from our estimate that portion of the total variation which is expected to be due to sampling error. Finally, we summarize the signal variance with a parsimonious time series model in which each performance measure consists of a persistent component and an idiosyncratic, non-persistent component.

A. Setup and Notation

More formally, let $\hat{\delta}_{jt}$ be a 1x2 vector containing estimates of two different performance measures for a particular school (j=1,...,J) in a given year (t=1,...,T). Each of these estimates is derived from a student-level regression of the form:

(1) $Y_{ijt} = \delta_{jt} + X_{ijt}\beta + u_{ijt}$,

where Y_{ijt} would be a student performance measure (e.g. 5th grade math test) and X_{ijt} would

include any relevant student characteristics that are being controlled for. The key parameters of interest are the school-specific intercepts for each year (δ_{jt}). Without loss of generality, we will assume that X includes year dummies and the school-specific intercepts are normalized so as to be mean zero in each year. Thus, with no other covariates, these would simply represent the difference between the mean test scores for each school and the average test scores for all schools in each year.

Let δ_j be the 1x2T vector of school-specific intercepts from all years and for both performance measures. Estimation of equation (1) by standard fixed-effects methods yields unbiased estimates of these school-specific intercepts ($\hat{\delta}_j$) along with variance estimates for these parameters (S_j), where:

(2) $\hat{\delta}_j = \delta_j + \varepsilon_j$

and S_j is an estimate of the 2Tx2T variance matrix of the estimation error (ε_j). In other words, equation 2 states that school-specific estimates are composed of a signal component and a noise component, and the variance of the noise component is known.

Finally, we characterize the time-series properties of true school performance by assuming that school performance in each year (δ_{jt}) consists of two independent components: a persistent component (θ_{jt}) that is meant to capture differences across schools in curriculum, staff and facilities that would be expected to persist over time; and a non-persistent component (ψ_{jt}) that is meant to capture other idiosyncratic factors that do not disappear with sample size-- such as the weather on the day of the test or the presence of a particularly disruptive student in that

year. The persistent component is modeled as a 1st-order vector autoregression (VAR), with the two performance measures in each year depending on the previous years values of both measures plus new innovations that can be correlated across the measures. The non-persistent component is assumed to be independent across years, but allows for correlation in this component across performance measures in a given year. More specifically, we assume:

(3) $\delta_{jt} = \theta_{jt} + \psi_{jt}$,

where: ψ_{jt} is i.i.d. with $Var(\psi_{jt}) = \Psi$

and: $\theta_{jt} = \theta_{j,t-1}\Phi + \upsilon_{jt}$, where υ_{jt} is i.i.d. with $Var(\upsilon_{jt})=\Sigma$, and initially $Var(\theta_{j1})=\Gamma$.

The unknown parameters of this model are the variances and covariance of the non-persistent component (Ψ), the variances and covariance of the persistent component in the first year of the data (Γ) and the innovations to the persistent component (Σ), and the coefficients determining how lagged values of each performance measure influence current values (Φ).

B. Estimation

The ultimate goal is to develop linear predictions of each school's performance (δ_j) as a function of the observed performance estimates ($\hat{\delta}_j$). In general, the minimum mean squared error linear predictor is given by $\hat{\delta}_j\beta_j$ where $\beta_j = [E(\hat{\delta}_j'\hat{\delta}_j)]^{-1}E(\hat{\delta}_j'\delta_j)$. Moreover, if we assume normality in equations 2 and 3, then this estimator also gives the posterior mean ($E[\delta_j| \hat{\delta}_j]$) and is the optimal choice for any symmetric loss function. But this estimator depends

upon two unknown moment matrices -- $E(\hat{\delta}_j'\hat{\delta}_j)$ and $E(\hat{\delta}_j'\delta_j)$ -- which, based on equation 2,

can be further decomposed as follows:

$$(4) \quad E(\hat{\delta}_j'\hat{\delta}_j) = E(\delta_j'\delta_j) + E(\varepsilon_j'\varepsilon_j)$$

$$(5) \quad E(\hat{\delta}_j'\delta_j) = E(\delta_j'\delta_j)$$

Thus, constructing the optimal linear predictor requires estimates of the signal variance [$E(\delta_j'\delta_j)$] and the noise variance [$E(\varepsilon_j'\varepsilon_j)$] for each school. Similarly, to construct predictions of the persistent component in school performance (θ_j) requires estimates of the variance in the persistent component [$E(\theta_j'\theta_j)$]. Therefore, the first step in constructing the optimal linear predictor is estimating each of the variance components.

We estimate each of the variance components as follows. As mentioned above, estimation of equation 1 with the individual student test score data generates estimates of [$E(\varepsilon_j'\varepsilon_j)$] for each school-- namely S_j, the variance-covariance matrix for the school-specific intercepts. The variance in the signal and in the persistent component can each be calculated as a function of the parameters of the time-series model specified in equation 3, e.g. [$E(\delta_j'\delta_j)$] = $f(\Psi,\Phi,\Sigma,\Gamma)$. We estimate the time series parameters (Ψ, Σ, Γ and Φ, all 2x2 matrices, the first three symmetric) by noting that equation 4 implies that:

$$(6) \qquad E\left[\hat{\delta}_j'\hat{\delta}_j - S_j\right] = E\left[\delta_j'\delta_j\right] = f\left(\Psi,\Phi,\Sigma,\Gamma\right)$$

Thus, the time-series parameters can be estimated by Optimum Minimum Distance (OMD)

methods (Chamberlain, 1983), i.e. by choosing the parameters so that the theoretical moment matrix, f(Ψ,Φ,Σ,Γ), is as close as possible to the corresponding sample moments from the sample average of $\hat{\delta}_j' \hat{\delta}_j - S_j$. Let d_j be a vector of non-redundant (lower triangular) elements of $\hat{\delta}_j' \hat{\delta}_j - S_j$, and let g(Ψ,Φ,Σ,Γ) be a vector of the corresponding moments from the theoretical moment matrix. The OMD estimates of (Ψ,Φ,Σ,Γ) minimize the objective function:

$$(7) \qquad q = N\left[\bar{d} - g(\Psi, \Phi, \Sigma, \Gamma)\right]' V^{-1}\left[\bar{d} - g(\Psi, \Phi, \Sigma, \Gamma)\right]$$

where V is the sample covariance for d_j, and $\bar{d}$ is the sample mean of d_j. The value of the objective function (q) provides a measure of goodness of fit, and is distributed $\chi^2(p)$ if the model is correct, where p is the degree of over-identification (the difference between the number of elements in d and the number of parameters being estimated).

V. Results

A. Decomposing the Variation in 5th Grade Math and Reading Scores.

Table 3 reports the parameter estimates for fifth grade math and reading scores, both levels and gains, after having adjusted each student's level and gain score for race, gender and categories of reported parental education (using the same six categories reported in Table 1). In other words, all four series are intended to capture differences in school performance, after controlling for the more readily observable information on race, gender and parental education.

There are 4 findings worth noting in Table 3: First, after sorting out differences across schools that are due to sampling variation and other non-persistent differences in performance, the variation in school performance is small relative to the differences in student performance. Our estimate of the variance in adjusted school math test scores in 1994 (the variance in initial conditions, Γ) was just .061, implying a standard deviation of .247. Since test scores have been reported in student-level standard deviation units, this implies that the standard deviation in underlying differences in math performance at the school level is only one quarter of the standard deviation in 5th grade math test scores at the student level. There is even less of a signal to be found in reading test scores and in gain scores for both reading and math. Indeed, we estimate that the standard deviation in school reading score gains in fifth grade is only .077, one-thirteenth as large as the standard deviation in reading test scores among students.

Second, while we estimate that there is a considerable amount of persistence in math and reading performance over time at the school level (both levels and gains), we can strongly reject the hypothesis that school test scores are fixed effects. If school rankings were fixed over the long term, albeit with some short term fluctuations due to sampling variation and other non-persistent changes in school performance, we would expect the coefficients on lagged scores to approach one and for the variance in the persistent innovations to approach zero. Aside from sampling variation, the only innovations would be of the non-persistent type. As reported in Table 3, the coefficients on the lagged values range from .7 to .98, but generally range from .7 to .8. Moreover, the variance in persistent innovations-- changes in school performance that are at least partially passed on from one year to the next-- is estimated to be positive for all four outcomes. The p-values on the test of the fixed-effect specification reported in Table 3 suggest

that the fixed effect specification can be rather comfortably rejected.

Third, schools that perform well in 5th grade math also tend to perform well in 5th grade reading, in both levels and gains. This fact is reflected in the high correlations between reading and math in the initial conditions and persistent innovations. Even the non-persistent innovations are highly correlated for reading and math. This may not be surprising, given that elementary school teachers tend to teach both reading and math subjects. A disruptive student in the class or a particularly effective teacher would affect both math and reading test scores for a particular school. As a result, any differences in instructional quality are likely to be reflected in both.

Fourth, while the signal variance shrinks when moving from test score levels to gains-- as we might expect, as long as some of the differences in test score levels reflect the different starting points of students in the schools-- the variance in the non-persistent innovations in school performance does not. In fact, the non-persistent variance in math test scores doubles when moving from levels to gains, from .006 to .011. Recall that the variance in non-persistent innovations (Ψ) reflects the changes in student test scores that do not disappear with the number of students in the school, yet do not seem to be passed on from one year to the next. Some sources of such variation would be factors that would affect a whole school-- such as a dog barking in the parking lot on the day of the exam, a rainy day, an active construction site next door or transient factors affecting a single classroom-- one or two particularly disruptive students or strong "chemistry" between the teacher and the particular group of students. Because the gain scores use the difference in outcomes over two testing dates, such non-persistent variation would tend to be magnified (because there is no covariance between them). Meanwhile, it is the ratio of the variance in the signal to the total variance in school test scores that determines the degree to

which the raw data are misleading. That ratio falls when moving from test score levels to gain scores, both because the variance of the persistent signal shrinks and because the non-persistent variation rises.

Table 4 uses these estimates to decompose the variance in school mean test scores for students in the 5th grade in 1994 into 3 parts: that reflecting differences in performance that will at least partially be passed on from one year to the next (Γ), that reflecting sampling variation and that reflecting non-persistent variation that is independent from year to year (Ψ). Because the variance due to sampling variation is sensitive to the number of 5th grade students in a particular school, we report the proportions attributable to each source of variation by school size.

In both levels and gains, reading scores contain a lower ratio of signal to noise. For small schools (25 students per grade level), only 49 percent of the variance in reading levels and only 26 percent of the variance in reading gains reflect differences that will be partially passed on from one year to the next. This is due to the fact reported above, that there is little underlying variance in schools' performance in reading after adjusting for demographic differences between the students, at least relative to math performance. The remainder is due to either sampling variation or other types of non-persistent fluctuations in scores.

Moreover, for both reading and math, gain scores contain a much larger share of noise than test score levels. Sampling variation is not the reason. Rather, it is the rising importance of other types of non-persistent variation not sensitive to sample size that contributes to the noise in gain scores. Among small schools, the proportion of variance in reading test scores due to sampling variation increases only slightly from 44 to 51 percent between reading levels and reading gains. Among these same small schools, the proportion of variance in math performance

due to sampling variation actually declines from 26 to 23 percent in moving from levels to gains. However, the proportion of variance due to other types of non-persistent factors nearly triples from 8 percent to 22 percent when moving from reading levels to gains and more than triples from 7 percent to 25 percent in moving from math levels to math gains. Even in large elementary schools, with 100 students per grade level, less than half of the variance in reading gains (43%) and only about two-thirds of the variance in math gains (63%) is due to signal, with much of the remainder due to non-persistent fluctuations other than sampling error.

B. Predicting Performance Based upon Filtered Estimates

In focusing on a school's test score performance, parents and policymakers are not primarily interested in an unbiased estimate of a school's performance last year, but in using that evidence to draw some inferences about the state of educational quality in the current academic year and in future years. Table 5 compares the mean performance in 1999 for schools ranking in the top 10 percent in 5th grade math gains on two different measures: the simple means of math gains in 1997 and the filtered prediction that would have been made of a school's performance in 1999 using all of the data available through 1997. Thus, both predictions use only the data from 1997 or before. However, the filtered prediction incorporates information from reading scores and from prior years, and "reins in" the prediction according to the amount of sampling variation and non-persistent fluctuations in the data.

Table 5 reports the mean 1999 performance, cross-tabulated by whether or not the school was in the top 10 percent using the filtering technique and using the naive estimate based upon the actual 1997 scores. Sixty-five (65) schools were identified as being in the top 10 percent as

of 1997 using both the naive and the filtered predictions, and these schools scored .15 student level standard deviations higher than the mean school two years later in 1999. However, among the schools where the two methods disagreed, there were large differences in performance. For instance, among the 25 schools that the filtering method identified as being in the top 10 percent that were not in the top 10 percent on the 1997 actual scores, the average performance on 5th grade math gains was .124 student-level standard deviations above the average in 1999. On the other hand, among the 25 schools chosen using actual 1997 scores who were not chosen using the filtering technique, scores were .022 standard deviations *lower* than the average school in 1999. The next to last column and row in Table 5 reports the difference in mean scores moving across the first two columns or first two rows. Among those who were not identified as being in the top 10 percent by the filtering method, knowing that they were in the top 10 percent on the actual 1997 score provided very little information regarding test scores. In fact the test scores were -.006 standard deviations lower on average holding the filtered prediction constant. In contrast, among those were not identified as being in the top 10 percent on actual 1997 scores, knowing that they were selected using the filtering method was associated with a .140 standard deviation difference in performance. Apparently, the filtering method was much more successful in picking schools that were likely to perform well in 1999.

Moreover, the filtering technique provides a much more realistic expectation of the magnitude of the performance differences to expect. As reported in the last column of Table 5, the schools in the top 10 percent on the actual test in 1997 scored .453 standard deviations higher than the average school in 1997. If we had naively expected them to continue that performance, we would have been quite disappointed, since the actual difference in performance was only .115

standard deviations. On the other hand, among those who were chosen using the filtering method, we would have predicted that they would have scored .180 standard deviations higher than the average school in 1999 based upon their performance prior to 1998. The actual difference in performance for these schools was .160 standard deviations.

Table 6 compares the R^2 one would have obtained using 3 different methods to predict the 1998 and 1999 test scores of schools using only the information available prior to 1998. The first method is the "filtering method" described in the methodology section above. The second method is using the actual 1997 score and applying a coefficient of unity to it when predicting the 1998 and 1999 scores. The third method would be to use the 4-year average of math performance prior to 1998 (1994-1997) to predict 1998 and 1999.

Whether one is trying to anticipate math or reading, levels or gains in 5th grade, the filtering method leads to greater accuracy in prediction. The R^2 in predicting 5th grade math levels was .41 using the filtering method, .19 using the 1997 score and .29 using the 1994-97 average. The filtering method also calculates a weighted average using the 1994-97 scores, but it adjusts the weights according to sample size (attaching a larger weight to more recent scores for large schools) and uses both the math and reading score histories in predicting either. In so doing, it does much better than a simple average of test scores over 1994-97.

In predicting math or reading gain scores in 1998, the second column reports *negative* R^2 when using the 1997 scores alone. A negative R^2 implies that one would have had less squared error in prediction by completely ignoring the individual scores from 1997 score and simply predicting that performance in every school would be equal to the state average in 1998 scores. Of course, one could probably do even better by not ignoring the 1997 score, but simply applying

a coefficient of less than 1 to the 1997 score in predicting future scores. That is essentially what the filtering method does, while recognizing that the optimal coefficient on the 1997 score (and even earlier scores) will depend upon the amount of non-persistent noise in the indicator as well as the school size.

Although it performs better than either the 1997 score or the 1994-97 average in predicting 1998 and 1999 gains, the R^2 using the filtering method is only .16 on math gains and .04 on reading gains. This hardly seems to be cause for much celebration, until one realizes that even if the filtering method were completely accurate in predicting the persistent portion of school test scores, the R^2 would be less than 1 simply because a large share of the variation in school performance is due to sampling variation or other non-persistent types of variation. Because of these entirely unpredictable types of error, the highest R^2 one could have hoped for would have been .75 in predicting math levels, .60 in predicting reading levels, .55 for math gains and .35 in reading gains. For math gains, for instance, the filtering method was able to predict 16 percentage points of the 55 percentage points that one ever had a hope of predicting, implying an R^2 for the systematic portion of school test scores of .16/.55=.29.

One of the practical ways in which the improved prediction accuracy of the filtered estimates can help states is in identifying under-performing schools for interventions. As in many other states, the North Carolina school reform law provides the state Department of Public Instruction extraordinary powers to intervene in schools that perform poorly on the standardized test. In 1997, the state identified 15 schools with poor performance in both levels and gains and assigned "assistance teams" of 3-5 educators to work in these schools. The next year, all of the schools had improved enough to escape being designated as a "low-performing" school. In

summarizing the results of that first year, the state Department of Public Instruction claimed an important victory:

> "Last year, the assistance teams of 3-5 educators each worked in 15 schools, helping staff to align the instructional program with the Standard Course of Study, modeling and demonstrating effective instructional practices, coaching and mentoring teachers and locating additional resources for the schools. As a result of this assistance and extra help provided by local school systems, nearly all of these schools made exemplary growth this year and none are identified as low performing." (emphasis added)
>
> NC Department of Public Instruction, "ABCs Results Show Strong Growth in Student Achievement K-8", August 6, 1998

Indeed, the value of the assistance teams was lauded in *Education Week*'s annual summary of the progress of school reform efforts in the states.[5] However, given the amount of sampling variation and other non-persistent fluctuations in test score levels and gains, schools with particularly low test scores in one year would be expected to bounce back in subsequent years.

Figure 4 reports the trend over time in test scores for one of the fifteen schools singled out for assistance teams based upon their test scores in 1997. The solid line reports the actual gain scores in math and reading for 5th grade students in the school. The line with error bars reports our filtered estimate of the persistent part of school performance. The squares represent our estimate of the persistent part plus the non-persistent variation in test scores. In fact, we estimate that the school consistently had above average gains in math in 5th grade and roughly average gains in reading. Indeed, the poor performance in 1997 is estimated to have been due entirely to a transient decline in performance. It is hardly surprising that the 1997 performance was

---

[5] Kathleen Kennedy Manzo, "North Carolina: Seeing a Payoff" Education Week Vol. 18, No. 17, p. 165.

followed by a large improvement.

C. Test Score Improvements and Other Measures of Student Inputs

As the stakes have been raised for teachers and school administrators, many are concerned that teachers will substitute instruction on simple test-taking strategies for more valuable, but difficult-to-measure content. It is difficult to know how one would recognize the undesirable sort of "teaching to the test", even if one were able to sit in classrooms and observe for oneself. However, in this section, we explore how improvements in a school's performance on the math or reading tests in 5th grade were related to improvements along other dimensions-- such as student absences or the amount of homework students are reporting to be doing or the amount of TV that they are watching. Presumably, if classroom instruction were to improve in quality, the opportunity cost of missing a day in school would rise and absences would also decline. The implications are less clear for homework and TV watching-- if students were substituting effort outside of class for poor instruction in the classroom, an improvement in the quality of instruction could actually lead to less homework and more TV watching. Nonetheless, we were interested in learning how all three trends were related to the improvements in student performance.

Table 7 reports the correlations between each of 3 outcomes-- the proportion of students reporting 14 or more days absent, the proportion of students reporting doing 1 hour or less of homework per night and the proportion of students reporting to watch 6 or more hours of TV per day-- and math and reading levels and gains. The first row of Table 7 reports the correlation in initial conditions between each of these outcomes and math and reading levels in 5th grade.

While there seems to be very little relationship between a schools' mean math or reading performance and its rate of absence in 5th grade in 1994, there appears to be a strong negative relationship between test scores and the proportion of students reporting very little homework and the proportion watching a lot of TV, with correlations of -.4 to -.6. However, when a school improved its performance between 1994 and 1999, there appeared to be very little change in any of the three outcomes. Indeed, none of the correlations in the long-term changes were significantly different from zero.

A similar pattern is observed for gain scores, as reported in the bottom half of Table 7. Although there is some evidence of a negative correlation between gain scores in reading and math and excessive TV watching and reporting of little homework, there is no evidence that subsequent improvements in the systematic portion of test scores between 1994 and 1999 were associated with changes in these measures of student outcomes.

Interestingly, there is some evidence that the non-persistent fluctuations in test scores were related to the non-persistent fluctuations in the other outcomes. In other words, when test scores suddenly rose, this was often accompanied by a decline in excessive TV watching and the reporting of little home work, but these fluctuations did not seem to be sustained.

In sum, among the elementary schools that improved their performance between 1994 and 1999, there is little evidence that these indirect measures of student engagement-- days absent, homework or TV watching-- changed in any systematic way, despite the fact that these other outcomes were related to test scores in the base year. This need not mean that teachers were teaching to the test, but there is little evidence of any dramatic change in student involvement in school in the schools where test performance improved or deteriorated.

D. How Important are Classroom-level differences – Is it Schools or Teachers?

Hanushek, Kain and Rivkin (1998) interpret the lack of correlation in math gains across grades in the same school as providing evidence of the lack of correlation in teacher effects within schools, since different teachers typically do not teach multiple grades in elementary school. The ideal data set for resolving the issue would allow one to track the performance of different groups of students assigned to the same teachers. Neither our data, nor the data from Texas used by Hanushek, Kain and Rivkin (1998) would allow one to do so. However, there are five indirect ways of attempting to resolve the question that we explore below. In the end, we conclude that while differences in gains at the classroom level are important, we estimate that classroom differences account for only about half the difference between schools, rather than virtually all of the difference, as Hanushek, Kain and Rivkin (1998) would conclude.

1. The Relationship Between a School's Performance in 4th and 5th Grades

As in Hanushek, Kain and Rivkin (1998), we can use our methods to investigate whether the schools with above average-gain scores in 4th grade also experience above average gain scores in 5th grade, as a way of indirectly measuring the correlation in instructional quality within a school. Moreover, by explicitly accounting for the noise in these estimates, our approach addresses concerns that the Hanushek et al. result was biased toward finding no association. The correlations of unfiltered estimates would be biased toward zero. As a result, it may not be surprising that one sees little correlation in a school's 4th and 5th grade performance. Studying two different grade levels within the same school over time complicates the model described above, since the sample of students being tested each year cannot be treated

as independent-- the 4th grade students in year t are the same sample of students in 5th grade in year t+1, and one might expect a correlation in each cohorts performance between 4th and 5th grade. Therefore, we adapted our model to allow for such cohort effects, which we incorporate as a correlation between the non-persistent component of the test score from 4th grade in year t and 5th grade in year t+1.[6]

Table 8 reports the parameters for the model describing contemporaneous relationship between 4th and 5th grade math performance in schools over time. Many of the results reported in Table 8 are consistent with the results reported earlier-- a relatively high degree of persistence from one year to the next in both math levels and math gains in a given grade, much less signal variance in math gains than in math levels, a substantial degree of non-persistent fluctuation in math performance, the variance of which essentially doubles when moving from test score levels to test score gains. However, there are 3 important results contained in Table 8 that are not reported in the earlier tables:

First, the amount of year-to-year fluctuation in performance attributable to the passage of successive cohorts of children is substantial. Cohorts that do exceptionally well on 4th grade math levels also tend to do well on 5th grade math levels (a correlation of .75). (This is the i.i.d. component alone. Much of the sampling variation is also likely to follow a cohort, since for any given sample, the 4th and 5th grade scores will be correlated.) The results in Table 8 suggest that 88 percent of the i.i.d. component in 4th grade math scores and 63 percent of the 5th grade i.i.d. variance is due to a shared cohort component. Recall that the i.i.d. component reflects

[6]This also required merging the test performance of individual students across 3 years rather than just across 2 years.

idiosyncracies of a particular cohort of students that does not disappear with sample size. Potential sources would include a dog barking in the parking lot of the school on the day of the test or one particularly disruptive student in a class or the "chemistry" between a teacher and a particular group of students. Since much of the i.i.d. component seems to be passed on when a cohort moves from 4th to 5th grade, one might conclude that the latter two types of "noise" account for a large share of such volatility. Moreover, cohorts that do exceptionally well in their 4th grade math gains tend to fair slightly less well in 5th grade math gains (a correlation of -.20).

Second, schools with strong *levels* of math performance in 4th grade tend to do well in math in 5th grade. Indeed, both the correlation in initial conditions and the correlation in the contemporaneous innovations is a positive .7. However, consistent with Hanushak et al., we find no evidence that the schools that have substantial *gains* in math performance during 4th grade also have substantial gains in math performance in 5th grade. In fact, there is a small negative correlation in math gains in 4th and 5th grades in initial conditions and zero correlation in subsequent innovations of the type that is partially passed on from one year to the next. These are contemporaneous correlations, and do not reflect the cohort effects that are accounted for elsewhere.

One interpretation is that the tests only measure a subset of skills, and that schools that emphasize certain math skills in 4th grade have little room to gain in 5th grade. An alternative interpretation is that there little correlation in the quality of instruction in 4th and 5th grades within a school-- that it is teachers that matter and that the quality of teaching is only weakly correlated within schools.

One admittedly puzzling result is the negative correlation in the contemporaneous, non-

persistent variation in 4th and 5th grade math levels. To the extent that the contemporaneous shocks are likely to be shared-- for example, the dog-barking outside the school on test day or the rainy weather on test days-- we would have expected a positive relationship in the contemporaneous shocks to test score levels, similar to that reported for test score gains. We would note that there is relatively little variance in the non-persistent component for either 4th or 5th grade math levels and so even a very small negative covariance estimate is likely to lead to a substantial negative correlation. Nonetheless, it may also reflect some error in our specification of the error structure. Both the levels and gains specifications in Table 8 fail the goodness of fit tests at the .001 and .003 level respectively.

2. The Relationship Between School Size and Variation in Performance

An alternative test of the importance of classroom effects is based on the relationship between the degree of variance in student gains across schools-- in initial conditions, persistent innovations and non-persistent innovations-- and the number of classrooms in each school. If there were substantial differences in instructional quality within schools, then the law of averages would imply that the more classrooms a school had for a given grade-level the less heterogeneity we would expect to see in initial performance and in the persistent and non-persistent fluctuations in performance. To shed some light on this hypothesis, we estimated the specification in the third column of Table 3, separately for schools with varying numbers of classrooms in the 5th grade. Those results are reported in Table 9.

All three sources of variance tend to shrink as the number of classrooms increases. For instance, schools that had an average of 0 to 2.5 classrooms in the 5th grade over the time period

we studied, had an initial variance in reading gains of .008. The initial variance for schools for twice the number of classrooms (2.5 to 4.5) was roughly half as large (.005) and the initial variance with four times as many classrooms (over 4.5) was roughly one quarter as large (.002). The variance in the persistent and non-persistent innovations also declines with the number of classrooms.

Admittedly, there may be other explanations, such as that the large schools are more vigilant in seeking uniformity in curricula for different classrooms in the same grade. However, if there were a random teacher component influencing student gains within schools, we would expect each of these variances to decline for larger schools. The results in Table 9 are consistent with such an explanation.

3. Decomposing School and Classroom Effects for One Year

A third test for the importance of individual teacher differences is to directly estimate the amount of within- and between-school heterogeneity in classroom gains for a single year. Although the NC data do not allow us to track a given teacher over time, we do know which students shared the same classroom in 1996. Rather than having observations for the same school over multiple years, we had observations from different classrooms within each school in one year. Just as we decomposed variance in school performance over time into a persistent school effect and independent fluctuations over time, we were able to estimate a analogous specification, estimating fixed school effects and within school variance in gains at the classroom level. (With only one year of data, however, we are unable to decompose the school or classroom variance into persistent and non-persistent components.) Finally, because our

estimator was constructed to handle schools with a given number of years of data, we had to limit schools to a fixed number of classrooms. As a result, we limited ourselves to those schools with 3 or more classrooms in 5th grade in 1996. For the schools with more than 3 classrooms, we randomly selected 3 to use in our analysis.

The results of decomposing 5th grade math gains into a school and classroom component are reported in Table 10. There are two results worth noting. First, half of the variance in math gains at the classroom level is attributable to systematic differences between schools (.027), while the other half is attributable to differences in classroom gains within schools (.026). The same is true of reading gains in 5th grade (.006 is attributed to the school component, and .007 is attributed to the classroom component), although, as before, there is simply much less heterogeneity in reading gains across schools or classrooms. Second, the estimates are relatively stable across schools with different numbers of classrooms (particularly for the classroom component), in contrast to the smaller variance seen among larger schools in Table 9.

Together, these results suggest two conclusions: first, much of the variation in school performance is coming from within-school variation in classroom performance; second, the reason for observing less variation in performance across large schools is because the effects of individual teachers on student performance tend to average out in larger schools.

4. Comparing the Correlation in Math and Reading Gains in Elementary and Middle Schools

A fourth indirect way to test the importance of classroom-level differences in student gains is to compare the correlations between math and reading gains for students by grade. In elementary school, a student will typically have the same teacher for all subjects. Therefore,

elementary schools with particularly strong instruction will tend to have similar gains in both reading and math. In middle school, students generally will have different teachers in different subjects. And, to the extent that schools with strong math instructors need not have strong instructors in other subjects, we might expect to see less of a correlation in math and reading gains at the school level in middle schools.

Table 11 reports the correlations in initial conditions, in the change in the persistent component of school performance between 1994 and 1999 and in the non-persistent component for reading and math gain scores in various grades. For each component, the correlation between reading and math gains at the school level is smaller in middle school than in elementary school. For instance, the correlation in math and reading gains at the school level in 1994 is estimated to be .762 in the fourth grade, .737 in fifth grade, .506 in 7th and .598 in sixth. Similarly, the change in the persistent component between 1994 and 1999 in math and reading is highly correlated in fourth and fifth grade (.700 and .664 respectively) and is less correlated in 7th and 8th grade (.434 and .421).

5. Comparing the Pattern of Correlation Over Time to Teacher Turnover

A final indirect test of the importance of heterogeneity in instructional quality within schools is to compare the time-pattern of correlation in the persistent part of math and reading gains over time with what we know about the average teacher tenure. If the persistent school effects were simply the aggregation of teacher effects, and if schools in hiring new teachers were simply pulling a random draw from the pool of teacher quality, we would expect the correlation in a school's performance over time to closely reflect the proportion of teachers remaining after 1

year, after 2 years, etc. The pattern of correlation over time is actually surprisingly close to that pattern.

Figure 5 reports the time-pattern in correlation in the persistent component of math and reading 5th grade gain scores after 1, 2, 3, 4 and 5 years respectively (after accounting for sampling variation and the non-persistent fluctuations in scores). Figure 5 also reports estimates of the proportion of elementary school teachers in North Carolina that had been with their school for 1 or more years, 2 or more years, 3 or more years, etc. These estimates are derived from 908 respondents to the 1993-94 Schools and Staffing Survey who reported teaching in North Carolina (We thank Terry Moe for providing us with these estimates.) For instance, 85 percent of teachers are estimated to remain with their employer for 1 or more years, 75 percent for 2 or more years, 67 percent for 3 or more years, 57 percent for 4 or more years and 50 percent for 5 or more years. The correlation in a school's math and reading gains over time fades too, and only slightly more rapidly. For instance, after 5 years, the correlation in 5th grade math gains was roughly .3 while the correlation in 5th grade reading gains was roughly .35. While this evidence is far from conclusive, it suggests that one reason for the lack of strong correlation in school performance over time may be the high rate of teacher turnover.

Summary: Teachers or Schools?

We provide two pieces of evidence which, when taken together, suggest that heterogeneity in teacher quality accounts for a large share of between-school differences in performance. First, math and reading gain scores are more correlated in grades 4 and 5 when teachers typically teach the same subject than they are in grades 7 and 8 when teachers are more

likely to specialize by subject. Second, the correlation in gain scores for a particular school and grade level fades out over time only slightly faster than the rate of teacher turnover.

However, roughly half of the variance in classroom level gains seems to be "between schools"rather than "within schools." We would not conclude that classroom-level gain scores within the same school are independent. Classrooms with particularly strong or weak gains do tend to be "clumped" together by school. As reported in Table 10, in 5th grade, about half of the variance in classroom-level gains is between-schools rather than within-schools. This fact is hard to see when comparing gain scores at different grade levels. There are two reasons for this: First, sampling variation and other non-persistent variation in 4th and 5th grade scores bias any correlation in unfiltered scores toward zero. Second, after accounting for both such sources of statistical noise, we estimate that there is actually a negative correlation in math gain scores in 4th grade with math gain scores in 5th grade. Presumably, the negative correlation in gain scores across grades within schools is due to the fact that standardized tests focus on only a finite number of skills. Some teachers present a tough act (or easy act) to follow, at least on the limited set of skills that are being tested. We would infer that Hanushek, Kain and Rivkin (1998) found a zero correlation in gain scores in adjacent grades primarily because statistical noise obscured a negative correlation.

VII. Conclusion

Wall Street does not rely upon short-term fluctuations in quarterly income as the sole metric for evaluating changes in a public corporations' financial health. Rather, they tend to downweight those fluctuations, depending upon what they know about the volatility of a particular company's earnings in the past and the seasonal patterns of earnings for firms in that industry. Market watchers use what they know about past volatility to put the most recent news into context.

In this paper, we implement that insight in a systematic way in the case of school report card measures. Parents and school administrators are only interested in recent test performance to the extent that it tells them something about current and future levels of performance. By focusing on the predictive usefulness of past performance, the technique we propose conveys a much more accurate impression of a schools' performance and presents a much more sound foundation upon which to build an accountability system.

There are costs to employing such a technique to characterize a schools' progress. First, given the limited amount of information conveyed in any single year of test scores, a school may have to show several years of sustained improvements before our summary measure would register a significant improvement. Second, for similar reasons, it would take longer for small schools to muster enough evidence to warrant an improvement in their ranking. However, these costs must be weighed against the cost of volatility on the incentives for teachers and school administrators.

The estimation technique we use to decompose the variance in school-level test scores also yielded a number of substantive implications. First, one must be cautious in using gain

scores in an accountability framework, whether one is evaluating schools or teachers. There is much less signal variation and relatively more variation due to non-persistent factors in gain scores than in test score levels. Moreover, the gain any teacher is likely to achieve with his or her students seems to depend upon the quality of instruction provided in the previous year. Large (or small) gains one year tend to be followed by small (or large) gains in the next year. Also, the schools that achieve impressive gains in one grade may not achieve impressive gains in other grades. In other words, one should not evaluate a school based upon the gains in any particular grade level. Although gain scores are often touted as better indicators of "value-added" by a school, their usefulness will be quite limited without the filtering technique we propose.

Second, we found little evidence that schools with substantial improvements in test performance over time improved on any measures of student engagement. Although homework and TV watching were strongly related to math test score gains in 5th grade in the base year (1994), there was no evidence that the schools with the greatest improvements in performance after 1994 exhibited improvement on any of these other dimensions. Such results would be consistent with the hypothesis that schools began tailoring their curricula to improve performance on the tests, without generating similar improvements on other measures.

Finally, much indirect evidence points to the importance of heterogeneity in instructional quality at the classroom level accounting for a large share of the differences in gain scores across schools and within schools. Unfortunately, we were not able to test this hypothesis directly without being able to track the performance of different groups of students assigned to the same teachers over time. In future work, we hope to be able to attack the question more directly, by tracing trends in school as well as teacher performance.

References:

Chamberlain, Gary, "Panel Data" Chapter 22 in Zvi Grilliches and Michael D. Intrilligator (eds.) Handbook of Econometrics, Vol. II (New York: Elsevier Science, 1984). pp. 1247-1318.

Coleman, James S., E.Q. Campbell, C.J. Hopson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York *Equality of Educational Opportunity* (Washington, DC: U.S. Department of Health, Education and Welfare, 1966).

Hyslop, Dean and Guido W. Imbens "Bias from Classical and Other Forms of Measurement Error" National Bureau of Economic Research Working Paper No. T0257, August 2000.

McClellan, Mark and Douglas Staiger, "The Quality of Health Care Providers" National Bureau of Economic Research Working Paper No. 7327, August 1999.

Morris, Carl. "Parametric Emprical Bayes Inference: Theory and Applications" Journal of the American Statistical Association, (1983) Volume 381, No. 78, pp. 47-55.

Normand, Sharon-Lise, Mark Glickman and Constantine Gastonis, "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications" Journal of the American Statistical Association (1997) Vol. 92, No. 439, pp. 803-814.

Olson, Lynn "The Push for Accountability Gathers Steam" Education Week, February 11, 1998.

Rivkin, Steven, Eric Hanushek and John Kain, "Teachers, Schools and Academic Achievement" National Bureau of Economic Research Working Paper No. 6691, August 1998. (Revised, April 2000)

Sandham, Jessica L. "Florida OKs First Statewide Voucher Plan" Education Week, May 5, 1999.

Tarcy, Brian "Town's Scores the Most Improved" *Boston Globe*, December 8, 1999, p. C2.

Table 1.
Characteristics of the Matched and Non-Matched Sample of 4th and 5th Grade Students in 1999

	Non-Matched	Matched
% of 4th and 5th Grade Students	34.2	65.8
Mean Math Score	153.8	156.5
S.D. in Math Score	11.1	10.5
Mean Reading Score	150.5	152.4
S.D. in Reading Score	9.5	9.1
Percent Female	47.4%	50.1%
Percent Black	35.1	27.7
Percent Hispanic	5.4	2.2
Parental Education:		
H.S. Dropout	16.6%	9.8%
H.S. Graduate	47.1	43.7
Trade/Business School	4.6	5.3
Community College	11.3	14.2
Four-Year College	16.5	21.9
Graduate School	3.9	5.1
Sample Size	69,388	133,305

Note: Each of the differences above were statistically significant at the .05 level.

Table 2.
Proportion Ranking in the Top 10 Percent
on 5th Grade Test Scores 1994-1999

Number of Years in Top 10% during 1994-99	Adjusted Levels		Adjusted Gains		*Expected Proportion*	
	Math	Reading	Math	Reading	*Annual Lottery*	*Certainty*
Never	.6868	.6499	.6398	.6152	*.5314*	*.9000*
1 Year	.1633	.2181	.2237	.2383	*.3543*	*0*
2 Years	.0749	.0727	.0694	.0940	*.0984*	*0*
3 Years	.0369	.0235	.0380	.0336	*.0146*	*0*
4 Years	.0190	.0179	.0213	.0179	*.0012*	*0*
5 Years	.0101	.0089	.0045	.0011	*.0005*	
All 6 Years	.0089	.0089	.0034	0	*.000001*	*.1000*

Note: Test scores were adjusted for the race, parental education and gender of the students and then averaged by grade level within schools.

Table 3.
Estimates of Parameters Describing Time Series of School Effects for 5th Grade Math and Reading

	Adjusted Levels		Adjusted Gains	
	Math	Reading	Math	Reading
Coeff on $Math_{t-1}$ $\Phi_{1.}$	.694 (.028)	-.075 (.020)	.767 (.046)	.006 (.025)
Coeff on $Read_{t-1}$ $\Phi_{2.}$	.255 (.039)	.989 (.031)	.046 (.102)	.780 (.070)
Variance in Initial Conditions (Γ) *[Implied Standard Dev.]*	.061 (.004) *[.247]*	.030 (.002) *[.174]*	.023 (.002) *[.152]*	.006 (.001) *[.077]*
Correlation in Initial Conditions	.796 (.018)		.737 (.051)	
Variance in Persistent Innovations (Σ) [Implied Standard Dev.]	.018 (.002) *[.135]*	.005 (.001) *[.072]*	.010 (.001) *[.100]*	.0016 (.0005) *[.040]*
Correlation in Persistent Innovations	.673 (.049)		.623 (.104)	
Variance in Non-persistent Innovations (Ψ) [Implied Standard Dev.]	.006 (.001) *[.080]*	.005 (.001) *[.069]*	.011 (.001) *[.106]*	.005 (.001) *[.071]*
Correlation in Non-persistent Innovations	.672 (.075)		.522 (.052)	
Test of Fixed-Effect Model (p-value)	.000		.000	
Over-identification Test (p-value)	.175		.009	
Number of schools	894		894	

Note: OMD estimates of parameters to describe time series in school effects $E(\delta_j'\delta_j)$. Scores are scaled so that a 1 unit change is equal to the unadjusted standard deviation in each score.

Table 4.
Decomposing the Variation in 1994 Math and Reading Test Scores

	Adjusted 5th Grade Levels		Adjusted 5th Grade Gains	
	Math	Reading	Math	Reading
Sample Size:	*Proportion Due to Persistent Differences Between Schools*			
25	0.662	0.488	0.517	0.263
50	0.760	0.623	0.584	0.355
100	0.829	0.724	0.625	0.429
Sample Size	*Proportion Due to Non-Persistent Differences Between Schools (Excluding Sampling Variation)*			
25	0.070	0.077	0.253	0.223
50	0.081	0.099	0.286	0.300
100	0.087	0.115	0.306	0.363
Sample Size	*Proportion Due to Sampling Variation*			
25	0.268	0.435	0.229	0.513
50	0.155	0.278	0.130	0.345
100	0.084	0.161	0.069	0.209

Table 5.
Performance of Schools in 1999 Identified as Being in the "Top 10%" in 1997
Based on Actual and Filtered Test Scores

5th Grade Math Gains

		Based on actual 1997 Score				
		School not in Top 10%	School is in Top 10%	Row Total	**Difference between Top 10% and the rest**	**Expected difference**
Based on filtered prediction of 1999 Score (from 1997)	School not in Top 10%	-0.016 (0.007) [N=779]	-0.022 (0.066) [N=25]	-0.016 (0.007) [N=804]	**-0.006** **(0.043)**	**0.385** **(0.034)**
	School is in Top 10%	0.124 (0.050) [N=25]	0.151 (0.026) [N=65]	0.144 (0.023) [N=90]	**0.027** **(0.052)**	**0.236** **(0.036)**
	Column Total	-0.012 (0.007) [N=804]	0.103 (0.027) [N=90]	0 (0) [N=894]	**0.115** **(0.024)**	**0.453** **(0.019)**
	Difference between top 10% and the rest	**0.140** **(0.042)**	**0.173** **(0.059)**	**0.160** **(0.023)**		
	Expected difference	**0.147** **(0.013)**	**0.095** **(0.012)**	**0.180** **(0.007)**		

Notes: Within the box, the entries report the mean of the 5th grade math gain score in 1999, along with standard errors of these estimates and the sample size in each cell. The columns of the table use actual scores in 1997 to assign schools to "top 10%" and to calculate the expected difference between the top 10% and the rest. The rows of the table use filtered predictions of 1999 scores, based only on data from 1994-1997, to assign schools to "top 10%".

Table 6.
Comparing the Accuracy of Alternative Forecasts
of 1998 and 1999 Test Scores

Test Score Being Predicted:	Unweighted R^2 when Forecasting 1998 and 1999 Scores under Alternative Uses of 1993-97 Data					
	Predicting Scores in 1998 (1-year ahead forecast R^2)			Predicting Scores in 1999 (2-year ahead forecast R^2)		
	"Filtered" Prediction	1997 Score	Average Score 1994-97	"Filtered" Prediction	1997 Score	Average Score 1994-97
Adjusted Score						
5th Grade Math	0.41	0.19	0.29	0.27	-0.02	0.13
5th Grade Reading	0.39	0.13	0.33	0.31	-0.05	0.24
Gain Score						
5th Grade Math	0.16	-0.27	0.09	0.12	-0.42	-0.01
5th Grade Reading	0.04	-0.93	-0.12	0.04	-0.85	-0.20

Note: The "filtered" prediction is an out-of-sample prediction, generated using only the 1993-1997 data.

Table 7. Estimated Correlations Between 5th Grade Test Scores and Other Outcomes (Derived From GMM Estimates)

	Proportion of students with 14 or more days absent		Proportion of students doing 1 hour or less of homework per day		Proportion of students watching more than 6 hours of TV per day	
	Math	Reading	Math	Reading	Math	Reading
Adjusted Levels						
Correlation between persistent component in 1994	-0.081 (0.056)	-0.098 (0.060)	-0.411 (0.049)	-0.410 (0.059)	-0.626 (0.091)	-0.480 (0.106)
Correlation between changes in persistent component, 1994-99[b]	-0.071 (0.099)	-0.242 (0.134)	-0.133 (0.084)	-0.148 (0.124)	–[a]	–[a]
Correlation between non-persistent components	-0.281 (0.144)	-0.218 (0.136)	-0.357 (0.092)	-0.287 (0.105)	-0.035 (0.129)	-0.244 (0.171)
p-value for test of independence	0.007	0.046	0.000	0.000	0.000	0.000
Adjusted Gains						
Correlation between persistent component in 1994	-0.088 (0.070)	-0.015 (0.093)	-0.128 (0.060)	-0.171 (0.085)	-0.247 (0.103)	-0.287 (0.117)
Correlation between changes in persistent component, 1994-99[b]	0.003 (0.112)	-0.001 (0.184)	-0.060 (0.098)	0.081 (0.158)	–[a]	–[a]
Correlation between non-persistent components	0.005 (0.096)	-0.073 (0.087)	-0.229 (0.064)	-0.280 (0.087)	-0.029 (0.096)	-0.071 (0.091)
p-value for test of independence	0.567	0.999	0.207	0.032	0.005	0.069
Sample size	894		923		896	

[a] Not reported because the persistent component of TV watching did not vary over time.
[b] The data for homework was only reported in a consistent way from 1995 through 1999.

Table 8.
Estimates of Parameters Describing Time Series of School Effects for 4th and 5th Grade Math (with Cohort Effects)

	Adjusted Math Levels		Adjusted Math Gains	
	4th	5th	4th	5th
Coeff on $4th_{t-1}$ $\Phi_{1.}$	.786 (.030)	-.040 (.030)	.792 (.031)	.001 (.030)
Coeff on $5th_{t-1}$ $\Phi_{2.}$	.064 (.026)	.873 (.028)	-.112 (.026)	.771 (.032)
Variance in Initial Conditions (Γ) *[Implied Standard Dev.]*	.049 (.003) *[.221]*	.057 (.004) *[.239]*	.020 (.002) *[.140]*	.023 (.002) *[.151]*
Correlation in Initial Conditions	.735 (.027)		-.181 (.075)	
Variance in Persistent Innovations (Σ) *[Implied Standard Dev.]*	.016 (.001) *[.125]*	.016 (.001) *[.127]*	.009 (.001) *[.093]*	.011 (.001) *[.106]*
Correlation in Persistent Innovations	.729 (.046)		-.013 (.120)	
Variance in Non-persistent Innovations (Ψ) *[Implied Standard Dev.]*	.005 (.001) *[.070]*	.007 (.001) *[.083]*	.012 (.001) *[.108]*	.010 (.001) *[.100]*
Correlation in Non-persistent Innovations	-.493 (.180)		.349 (.092)	
Correlation due to Cohort Effect (4th to 5th)	.748 (.124)		-.197 (.060)	
Over-identification Test (p-value)	.001		.003	
Number of schools	830		830	

Note: OMD estimates of parameters to describe time series in school effects $E(\delta_j'\delta_j)$. Scores are scaled so that a 1 unit change is equal to the unadjusted standard deviation in each score.

Table 9
Selected variance estimates for 5th grade gain scores
by average number of 5th grade classrooms

		Average number of 5th grade classrooms		
	Full Sample	0 to 2.5	2.51 to 4.5	Over 4.5
Variance of Initial Conditions in 1994				
Math Gain	0.023 (0.002)	0.037 (0.005)	0.019 (0.003)	0.014 (0.002)
Reading Gain	0.006 (0.001)	0.008 (0.002)	0.005 (0.001)	0.002 (0.001)
Variance of Change in Persistent Component, 1994-1999				
Math Gain	0.035 (0.003)	0.050 (0.006)	0.035 (0.004)	0.022 (0.003)
Reading Gain	0.007 (0.002)	0.006 (0.003)	0.008 (0.002)	0.002 (0.001)
Variance of Non-Persistent component				
Math Gain	0.011 (0.001)	0.012 (0.002)	0.010 (0.001)	0.005 (0.001)
Reading Gain	0.005 (0.001)	0.007 (0.001)	0.004 (0.001)	0.003 (0.0004)
Sample size	894	250	437	207

Table 10
Decomposition of 5th grade gain scores from 1996
into school and classroom components
(by average number of 5th grade classrooms)

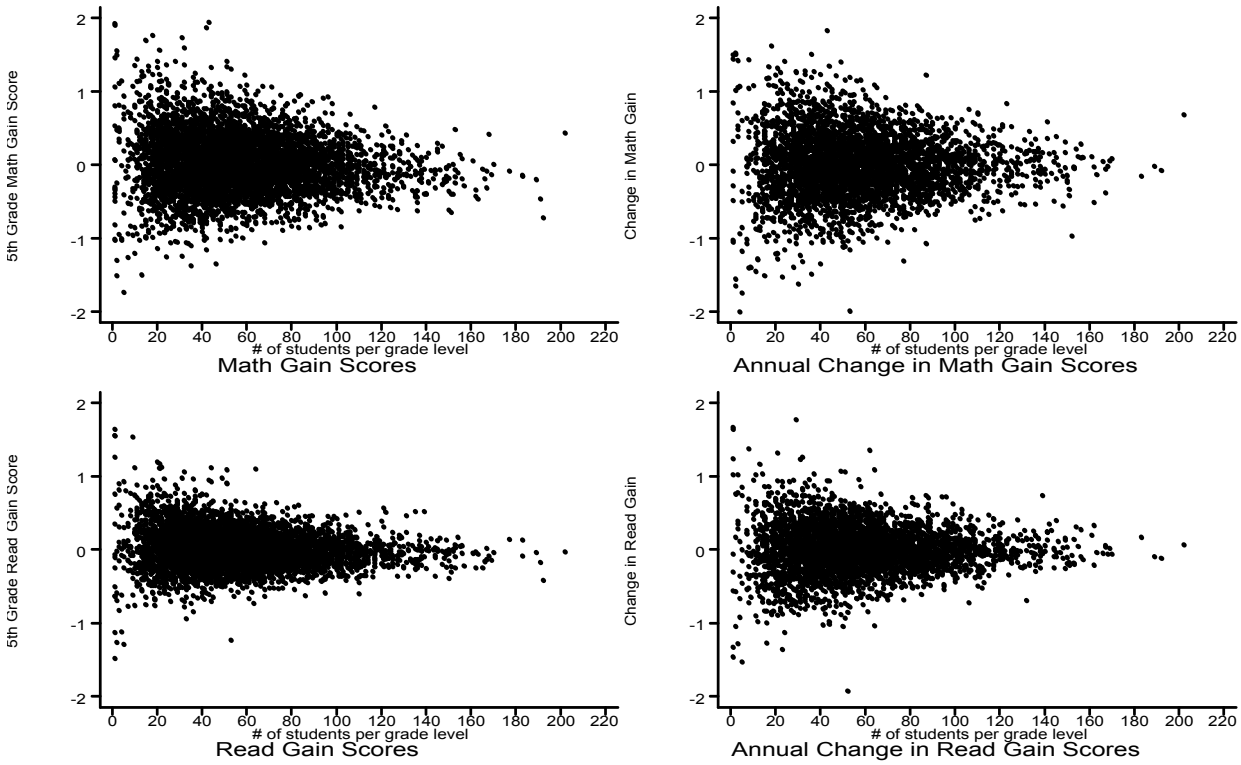
		Average number of 5th grade classrooms		
	Full Sample	0 to 3.5	3.51 to 5	Over 5
School Component				
Variance in Math Gain	0.027	0.031	0.030	0.016
[Implied Standard Dev.]	(0.002)	(0.004)	(0.004)	(0.003)
	[0.164]	*[0.177]*	*[0.174]*	*[0.128]*
Variance in Reading Gain	0.006	0.004	0.010	0.004
[Implied Standard Dev.]	(0.001)	(0.001)	(0.002)	(0.002)
	[0.080]	*[0.063]*	*[0.099]*	*[0.059]*
Correlation in Gains	0.654	0.740	0.617	0.708
	(0.051)	(0.108)	(0.061)	(0.143)
Classroom Component				
Variance in Math Gain	0.026	0.021	0.024	0.023
[Implied Standard Dev.]	(0.002)	(0.003)	(0.003)	(0.003)
	[0.161]	*[0.144]*	*[0.156]*	*[0.151]*
Variance in Reading Gain	0.007	0.005	0.004	0.007
[Implied Standard Dev.]	(0.001)	(0.002)	(0.001)	(0.002)
	[0.081]	*[0.074]*	*[0.066]*	*[0.081]*
Correlation in Gains	0.475	0.321	0.738	0.260
	(0.058)	(0.122)	(0.097)	(0.120)
Sample size (schools)	719	255	297	167

Note: Sample consists three randomly chosen classrooms from all schools with at least three 5th grade classrooms in 1996.

Table 11
Correlation Between Math and Reading Gain Scores
in Elementary School (common teacher) and Middle School (different teacher)

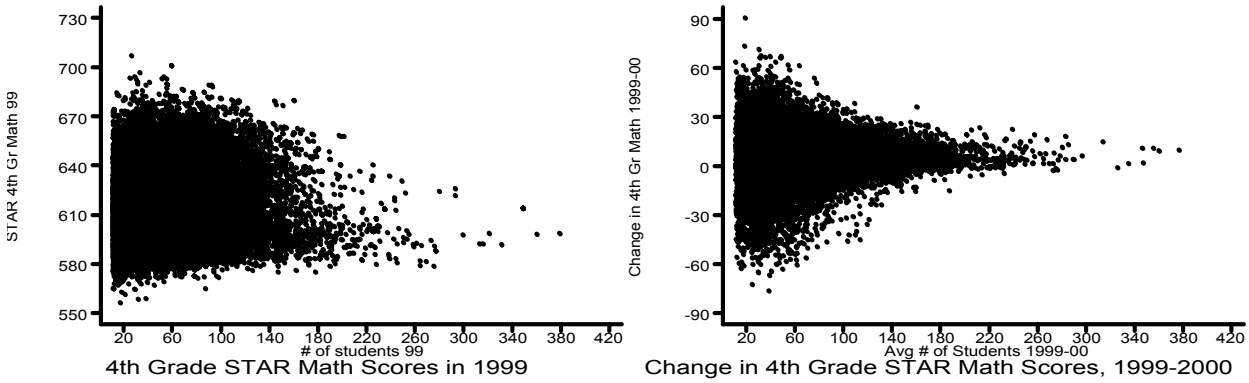
	Elementary School (same teacher in all subjects)		Middle School (different teachers by subject)	
	4th Grade	5th Grade	7th Grade	8th Grade
Correlation between math gain score and reading gain score in:				
Persistent component in 1994	0.762 (0.057)	0.737 (0.051)	0.506 (0.119)	0.598 (0.100)
Change in persistent component, 1994-99	0.700 (0.059)	0.664 (0.071)	0.434 (0.121)	0.421 (0.097)
Non-persistent component	0.579 (0.053)	0.522 (0.052)	0.184 (0.130)	0.490 (0.062)
Sample size (schools)	877	894	324	412

Figure 1



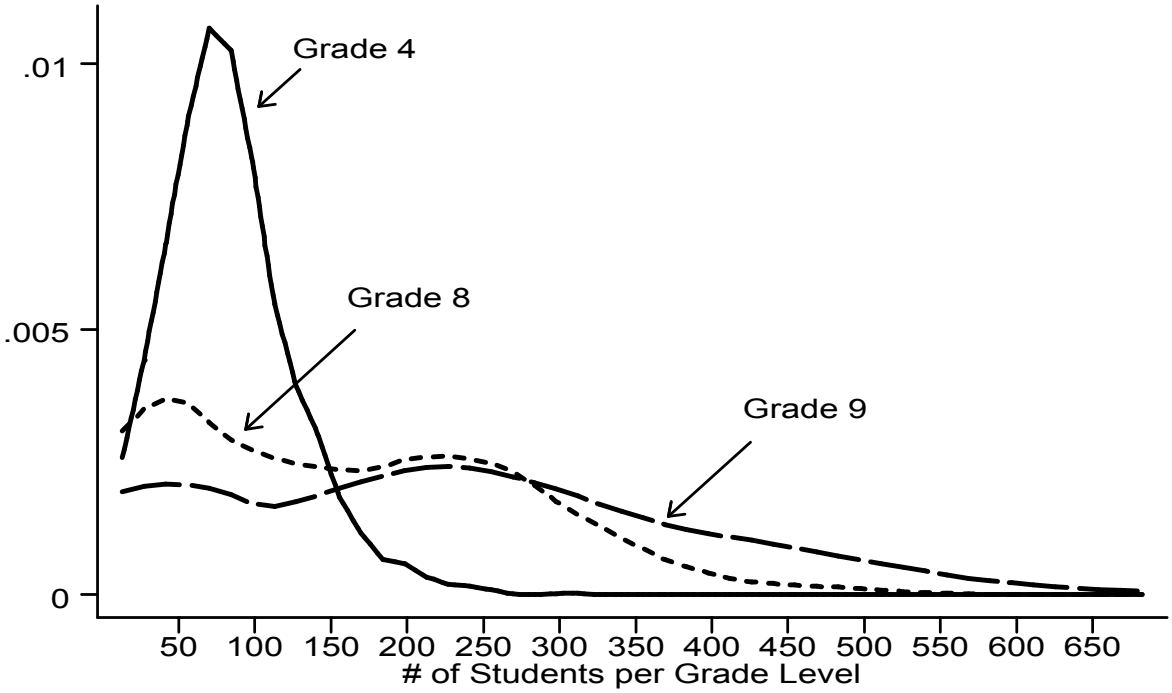
Math and Reading Gains and Changes by Sample Size

Figure 2.



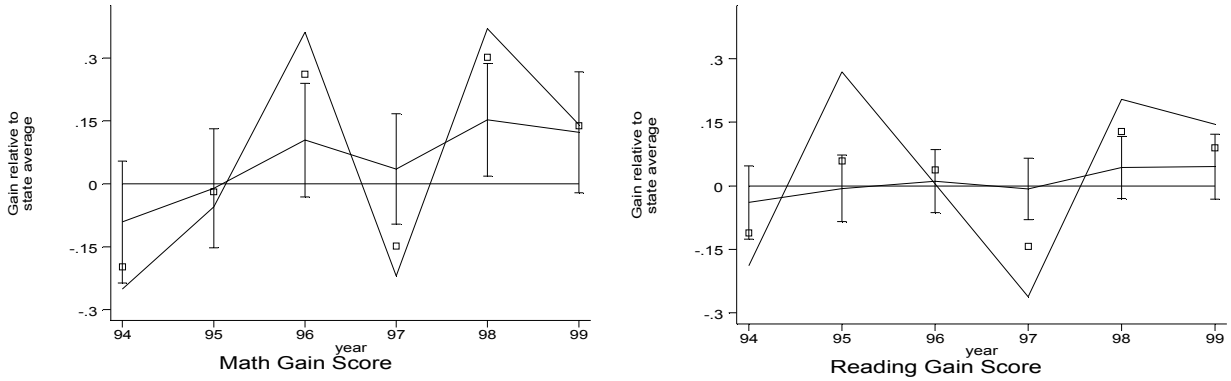
CA STAR Math Scores by School Size

Figure 3



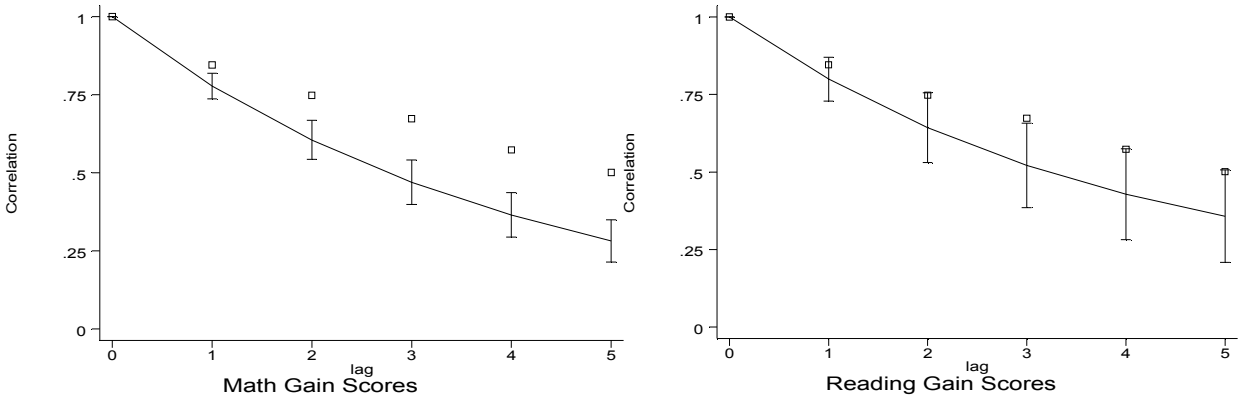
Distribution of School Size by Grade Level

Figure 4



Math and reading gain scores for 5th grade between 1994 and 1999 (line without error bars) compared to filtered estimates of the signal component (squares) and the persistent component (line with 90% confidence bars) for a school averaging 40 students per year, selected for North Carolina's Assistance program based on 1997 performance.

Figure 5



Correlation in math (left panel) and reading (right panel) gain scores between 1999 and previous years as estimated from the persistent component of the signal (line with 90% confidence bars) and as implied by teacher turnover rate of in North Carolina (squares)